sociological science

Supplement to:

Boutyline, Andrei. 2017. "Improving the Measurement of Shared Cultural Schemas with Correlational Class Analysis: Theory and Method." Sociological Science 4: 353-393.

APPENDIX A. CCA Algorithm

Correlational class analysis can be easily implemented in any programming environment which supports network partitioning by modularity maximization. It consists of four steps:

1. Create a matrix *G* of absolute row correlations between survey respondents.
2. Set statistically insignificant correlations to 0 to reduce noise (e.g., using *t*-tests[1]).
3. Import *G* into a network analysis package, treating it as an adjacency matrix.
4. Use the existing network partitioning routines to produce the class assignments.

In the *R* statistical environment with the *igraph 0.7* library, this can be implemented as:

```
CCA <- function (dataset, min.significant.row.cor = 0.60) {
   C <- abs(cor(t(dataset)))                    # 1st step
   C[C < min.significant.row.cor] <- 0          # 2nd step
   G <- graph.adjacency(C, mode="undirected",
             weighted = TRUE, diag = FALSE)     # 3rd step
   leading.eigenvector.community(G)$membership # 4th step
}
```

A more full-featured implementation of the method is available on *CRAN*, and can be installed in *R* with `install.packages("corclass")`.

See Appendix D for discussion of how to treat respondents with zero variance.

APPENDIX B. Simulating the Theorized Model

Procedure 1—Initial linear simulation procedure (fixed inversion probability)

Step 1 (parameters): I first randomly set the maximum ranges of various broad simulation parameters by drawing them from the uniform distribution: schema variance $\upsilon \sim U[0.3,3]$, noise variance $\epsilon \sim U[0,3]$, maximum shift $\delta \sim U\{0,\ldots,3\}$ maximum scaling $\gamma \sim U\{1,\ldots,3\}$, and number of schematic classes $c \sim U\{2,\ldots,6\}$.

[1] My exploratory results suggest that more stringent cutoffs may produce more accurate results as long as they are not so extreme as to turn some nodes into isolates. I used $\alpha = 0.05$ as the cutoff for the simulations reported above, and $\alpha = 0.01$ for the GSS analyses. A `min.significant.row.cor` of 0.60 approximates a *t*-test at $\alpha = 0.01$ for rows of 17 variables.

Step 2 (schemas): Then, for each of the c classes, I randomly generate a schema vector $\rho = [\rho_1, \dots \rho_{10}]$ by drawing from the Normal distribution, $\rho_i \sim N(\mu = 0,\ \sigma^2 = v)$, and rounding to the nearest integer. Any duplicate vectors are discarded, and new vectors generated in their place, until I have c unique vectors. Then, I randomly set the counts $n_1, \dots, n_c$ of respondents in each schematic class, $n_i \sim U\{100{,}101, \dots{,}500\}$.

Step 2b (range limits): The range of the 10 taste variables is then limited to $z_i = \pm[\max(|\rho_i|) * \gamma + \delta] \cap \mathbb{Z}$ (this limit is enforced at the end of Step 3).

Step 3 (responses): Finally, for each respondent $f \in [1, n = n_\rho]$ following schema ρ, I generate the 10-element response vector $X_f = (k_f * \rho) + \delta_f + \epsilon_f$ by first drawing the values of the vertical shift $\delta_f \sim U\{-\delta, \dots, \delta\}$ and the scaling and inversion factor $k_f \sim U\{[-\gamma, \dots, -1] \cup [1, \dots, \gamma]\}$. I then generate each respondent's noise vector ϵ_f by first determining *f*'s individual noise variance $\mathrm{E}_f \sim U(0, \epsilon)$, and then drawing each *i*th element $\epsilon_{fi} \sim N(\mu = 0, \sigma^2 = E_f)$, $i \in [1, \dots{,}10]$, rounded to the nearest integer. If any $X_{fi} \notin z_i$, where X_{fi} is the *i*th value of X_f, set it to the nearest value in z_i to enforce the range of the variable.

Procedure 2—Complete linear simulation procedure (random inversion probability)

To fully simulate the theorized model of schematic similarity as linear dependence, I extend Procedure 1 as follows:

Step 1: I now also draw a random inversion probability: $\zeta \sim U[0{,}0.5]$.

Step 3: I now draw a random inversion factor $z_f \in \{1, -1\}$, with $P(z_f = -1) = \zeta$. Since factor k_f now controls the scaling but not the inversion, I now restrict it to positive values: $k_f \sim U[1, \gamma]$. Each respondent f following schema ρ is generated by $X_f = (z_f * k_f * \rho) + \delta_f + \epsilon_f$.

Procedure 3: Polynomial functional form

I begin with Procedure 1 from Appendix B, omit step 2b, alter step 1, and replace step 3:
Step 1: I begin as before, though I no longer draw δ or γ. Instead, to determine the general form of the polynomial for this simulation, I first draw the count of polynomial terms $d \sim U\{1, \dots{,}10\}$. If $d \geq 2$, I then draw each exponent $D_i, i \in \{1, \dots, d\}$, from the set $\{0, \dots{,}10\}$, by sampling without replacement. If $d = 1$, I instead draw the sole exponent $D_1 \sim U\{1, \dots{,}10\}$. I then sort *D* in ascending order. Then, for each term in *D*, I generate a maximum scaling factor $\gamma_i \sim U[0{,}1]$.

Step 3: For each respondent $f \in [1, n_\rho]$ following schema ρ, I first I draw an exponent inversion factor $z_{fi} \sim \mathrm{U}\{1, -1\}$, and an exponent scaling factor $k_{fi} \sim U[1, \gamma_i]$ for each exponent D_i. I then generate the vector

$$X'_f = \mathbb{1}(D_1 = 0) * z_{f0} * k_{f0} + \sum_{i:i\in\{1,\ldots,d\},D_i\neq 0} [z_{fi} * \left(k_{fi} * \rho\right)^{D_i}],$$

where $\mathbb{1}$ is the indicator function. Since this procedure is prone to generate extremely wide ranges of variables that are not realistic for survey settings, before adding noise, I set all the values of X'_f that are below the 10th percentile or above the 90th percentile to equal the value at that percentile. Then I take the modified vector X'_f, linearly translate it to the [1,10] range, and round each element to the nearest integer. I then generate each element of their noise vector $\epsilon_{fi} \sim N(\mu = 0, \sigma^2 = \epsilon)$ for $i \in [1, \ldots, 10]$, rounding it to the nearest integer, and finally produce *f*'s response vector $X_f = X'_f + \epsilon_f$.

Procedure 4: Independent Subschemas

I begin with Procedure 2 from Appendix B, add step 1b after step 1, and replace step 3:

Step 1b (subschemas): For each schema ρ, I first set $i = 0$ and $j = 1$. Then, while $j \leq 4$ and $i < 10$, I draw a count of genres $m_j \sim U[1, 10 - i]$; assign genres $(i, i + m_j]$ to subschema *j*; increment i by m_j; and increment *j* by 1. If at the end of this procedure $j > 4$ and $i < 10$, I assign genres $(i, 10]$ to subschema $j - 1$, and decrement *j* by 1 so that *j* equals the number of different subschemas.

Step 3: I will use the notation ρ_{ij} to refer to the elements of schema *i* that were assigned to subschema *j*. For each respondent $f \in [1, n_\rho]$ following schema ρ, I first draw an individual noise variance $\mathrm{E}_f \sim U(0, \epsilon)$. Then, for each subschema $J \in [1, j]$, I simulate a partial response vector containing m_j cells, $X_{fJ} = \left(z_{fJ} * k_{fJ} * \rho_{pJ}\right) + \delta_{fJ} + \epsilon_{fJ}$, by drawing the values of the subschema vertical shift $\delta_{fJ} \sim U\{-\delta, \ldots, \delta\}$ the subschema scaling factor $k_{fJ} \sim U[1, \gamma]$, and the subschema inversion factor $z_{fJ} \in \{1, -1\}$ with $P\left(z_{fJ} = -1\right) = \zeta$. I then draw each element of the subschema noise vector $\epsilon_{fJi} \sim N(\mu = 0, \sigma^2 = \epsilon_f)$, $i \in [1, \ldots, m_j]$, rounded to the nearest integer. At the end of this step, I produce the respondent's vector X_f by concatenating together these *j* partial response vectors in the order they were generated.

Procedure 5: Multiway interactions

I again begin with Procedure 1. I then remove step 2b, alter steps 1 and, 2, and replace 3:

Step 1: I begin as before, though I omit drawing δ and γ. I instead draw the maximum number of terms per interaction $d \sim U\{1, \ldots, 4\}$, and the focal term weight $D \sim U\{1, \ldots, 4\}$.

Step 2: At the end of step 2, I now linearly translate all the schemas to the range $[1, \max(|\rho_i|) - \min(|\rho_i|) + 1]$.

Step 3: For each respondent $f \in [1, n_\rho]$ following schema ρ, I first draw an individual noise variance $\mathrm{E}_f \sim U(0, \epsilon)$. To determine which of *f*'s tastes interact with each other (*f*'s "interaction groups"), I first set $k = 0$ and $l = 1$. Then, while $k < 10$, I draw a count of genres $m \sim U[1, \min(10 - k, d)]$; assign genres $(k, k + m]$ to interaction group *l*; increment k by m; and increment *l* by 1. To generate each *i*th taste X_{fi} belonging to respondent *f*, I calculate a weighted geometric mean of tastes in the same interaction group. To do this, I begin with the *i*th element of ρ, exponentiate it to power *D*, multiply it with every *j*th element of ρ belonging to the same interaction group *m* as *i*, and finds the *k*th root of the product, where *k* equals $D - 1 + |m|$, and $|m|$ is the number of elements in *m*. I then generate X_{fi} by taking the result of this procedure and adding random noise $\epsilon_{fi} \sim N(\mu = 0, \sigma^2 = \mathrm{E}_f)$, rounded to the nearest integer.

APPENDIX C: Theory-Driven Changes to Pearson's Correlation Coefficient

Consider the absolute value of Pearson's correlation coefficient $|r(X, Y)| =$

$$\left| \frac{Cov(X, Y)}{\sqrt{Var[X]}\sqrt{Var[Y]}} \right| =$$

$$\left| \frac{E[(X - \bar{X})(Y - \bar{Y})]}{\sqrt{E[(X - \bar{X})^2]} * \sqrt{E[(Y - \bar{Y})^2]}} \right|$$

Different components of this formula make the coefficient invariant to inversion, scaling, and shift in the vector. Each of the three "sub-linear" scenarios I described earlier can be specifically accommodated by altering the relevant component (and rescaling the result to the $[0,1]$ range if needed). While I leave a fuller methodological treatment of this topic for future work, I derive some basic formulas below as an example of this approach.

No inversion. Most obviously, $|r(X,Y)|$ is invariant to inversion because of the absolute value operator. If inversion is to be interpreted as maximum schematic difference rather than schematic similarity, i.e., $r_{\sim}(X,Y) = 1 \rightarrow r_{\sim}(X,-Y) = 0$, the absolute value operator can simply be removed, with the resulting formula shifted and rescaled to [0,1]:

$$r_{\sim}(X,Y) = 0.5 * \left(\frac{Cov(X,Y)}{\sqrt{Var[X]}\sqrt{Var[Y]}} + 1 \right)$$

No scaling. To create a version of the coefficient that is sensitive to scaling, it is useful to note that correlation between a variable X and its multiple $Y = kX$ equals 1 because both the numerator and denominator of $r(X,Y)$ scale along with *k*:

$$r(X,kX) = \frac{Cov(X,kX)}{\sqrt{Var[X]}\sqrt{Var[kX]}} = \frac{k * Var[X]}{\sqrt{Var[X]}\sqrt{k^2 Var[X]}} = \frac{k}{k} = 1$$

It is possible to transform this mechanism into one that penalizes differences in scaling in proportion to the multiplier *k*:

$$r_{\times}(X,Y) = \left| \frac{Cov(X,Y)}{\max(Var[X], Var[Y])} \right|$$

If the variances of X and *Y* are equal, $r_{\times}(X,Y) = |r(X,Y)|$. However, if $Var[X] > Var[Y]$, $r_{\times}(X,Y) = \left|\frac{Cov(X,Y)}{Var[X]}\right| = |r(X,Y)| * \frac{\sqrt{Var[Y]}}{\sqrt{Var[X]}} = |r(X,Y)| * \sigma_y/\sigma_x$. Thus, as desired, $r_{\times}(X,kX) = |\frac{1}{k}|$ if $|k| > 1$, and $|k|$ otherwise. More broadly, for any $X \neq Y$, $r_{\times}(X,Y)$ will penalize their correlation in proportion to the ratio of their standard deviations.

No shift. Finally, it is possible to modify the correlation coefficient to penalize vertical shifts by replacing the variances in the denominator with the variables' second moments around the grand mean $\bar{M} = 0.5(\bar{X} + \bar{Y})$, yielding

$$r_{+}(X,Y) = \left| \frac{Cov(X,Y)}{\sqrt{E[(X-\bar{M})^2]} * \sqrt{E[(Y-\bar{M})^2]}} \right|$$

Since $Var[X] = E[(X-\bar{X})^2] < E[(X-\bar{M})^2], \forall\, \bar{M} \neq \bar{X} \in R$, this quantity will equal $|r(X,Y)|$ only if the two variables have the same mean, and will otherwise penalize it towards 0.

APPENDIX D: Zero-variance responses

Since correlation is normalized by the product of variances, it is undefined when the variance of either respondent is at absolute zero. The minimal implementation in

Appendix A requires that such respondents be dropped from the analysis. Since zero-variance respondents are relatively rare in empirical survey data (e.g., out of the 1532 respondents to the 1993 GSS musical tastes module, there was a total of only 4 with zero variance), dropping them will often be the most pragmatic solution. There are, however, empirical settings where zero-variance respondents can be common, such as when a group of respondents is reliably drawn to one extreme of the scale over the other (e.g., party-line voters on ballots). In such situations, this property of correlation can become a serious limitation necessitating a more considered solution.

The theory of cultural schemas does not appear to suggest any clear way of dealing with zero-variance respondents. Since such respondents express the same attitude towards all musical styles, their tastes literally contain no distinctions between any pair of genres. This lack of cultural judgment means that they are, in a sense, sitting out the game of distinction altogether. Thus, depending on theoretical context, an "undefined" schematic class membership may be justified. On the other hand, this lack of distinctions could be interpreted as a kind of "null schema" that specifies no contrasts between any genres. Following this logic, correlations between two zero-variance respondents could be set to 1, and their correlations with others to 0. (To keep CCA's accuracy directly comparable to RCA's, the algorithm I use to analyze the simulated datasets took this latter approach.)

In other theoretical settings, other solutions to the classification problem may be preferable. For example, in the absence of substantial differences in variance between respondents (e.g., when the survey items feature small numbers of response categories), it may be possible to use covariance instead of correlation, since the covariance between two zero-variance respondents equals zero. However, as long as multiplicative scaling ($Y = kX$) is one of the theorized schematic transformations, many intuitively appealing solutions may prove conceptually problematic. For example, in some theoretical settings, it may make sense to treat zero-variance responses as extreme cases of a low-variance schema, and thus to assign them to the schematic class containing respondents with the lowest average variance. But variance, unlike correlation, is not invariant to multiplication: if $Y = kX$, $Var[Y] = k^2 * Var[X]$. Therefore, in the presence of multiplicative scaling and a sufficiently broad range of *k*, any non-null schema could produce both low- and high-variance responses. This would then make "amount of variance" a problematic way of distinguishing between schemas.